

DOCUMENT RESUME

ED 150 171

TM 006 875

AUTHOR Bradshaw, Charles W., Jr.
TITLE The Use of Invariant Item Parameters to Derive an Absolute Score Metric.
PUB DATE [68]
NOTE 12p.
EDRS PRICE MF-\$0.83 HC-\$1.67 Plus Postage.
DESCRIPTORS *Equated Scores; *Item Analysis; *Mathematical Models; Multiple Choice Tests; *Scoring; Test Interpretation; Test Items
IDENTIFIERS Latent Trait Theory; Maximum Likelihood Estimation; Three Parameter Model

ABSTRACT

A method for determining invariant item parameters is presented, along with a scheme for obtaining test scores which are interpretable in terms of a common metric. The method assumes a unidimensional latent trait and uses a three parameter normal ogive model. The assumptions of the model are explored, and the methods for calculating the proposed statistics are appended. (Author/CTM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED150171

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CHARLES W.
BRADSHAW, JR.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

THE USE OF INVARIANT ITEM PARAMETERS TO

DERIVE AN ABSOLUTE SCORE METRIC

by

Charles W. Bradshaw, Jr.

TM006 875

ABSTRACT

This report points out the inadequacies of conventional test item parameters and the unsophisticated scoring systems currently used in mental testing programs. It is submitted that the variable nature of traditional item parameters precludes the possibility of establishing a stable score metric for comparing scores across test administrations. This paper presents a method for determining invariant item parameters along with a scheme for using them to arrive at test scores which are interpretable in terms of a common metric. The body of the report presents the model and explores its assumptions. A mathematical appendix is included which explains methods of calculating the proposed test item statistics and scores for the special case of multiple-choice dichotomous test items where omitted responses are treated as incorrect.

Key words: latent trait theory, normal ogive model, maximum likelihood ability estimates.

THE USE OF INVARIANT ITEM PARAMETERS TO DERIVE AN ABSOLUTE SCORE METRIC

by

Charles W. Bradshaw, Jr.

PROBLEM RECOGNITION

Scores on conventional multiple choice tests cannot currently be consistently compared across subsequent administrations which, a fortiori, precludes the possibility of deriving meaningful information from trend tracking procedures and techniques. The reason that comparisons between tests lack meaning is that the tests do not share a single common metric. A definite need exists in many testing programs for a common "yard stick" by which examinee, and group performance, can be compared from one administration to another.

PROBLEM SPECIFICATION

Why are conventionally scored tests not interpretable in terms of a common metric? The reason for this is related to the fact that item characteristics are not presently used to arrive at test scores for individuals. Ignoring the item characteristics when scoring individuals is a waste of valuable information.

Why have item characteristics been neglected in the past when scoring tests? This is primarily because the traditional item parameters derived from test items have not been "characteristic" in the true sense of the word. As a matter of fact, conventional item parameters vary from test group to test group. In order for item parameters to be truly characteristic, they must be invariant. Conventional item parameters are an artifact of the particular test population from which they were derived. For this reason, they do not lend themselves to construction of a stable metric for comparing test scores. Conversely, if invariant item parameters could be obtained, it would be possible to devise a common, stable metric for comparing scores across subsequent test administrations. It is the purpose of this paper to present a method for determining such invariant item parameters along with a scheme for using them to arrive at test scores which are interpretable in terms of a common metric.

SOLUTION*

Individual psychological traits can be characterized as manifest (observable) or latent (unobservable). Latent trait theory contends

*Many writers have espoused the use of the normal ogive function as a means of representing a theoretical mental test model. The author of this paper was most greatly influenced by the work of Lord and Novick (1968). For this reason, much of the material in this section will be found to closely parallel this classic work.

that manifest variables depend upon, or are functions of, underlying latent variables,

i.e., Manifest Variables = $f(\text{Latent Variables})$

In mental testing, the manifest variables are test scores and the latent variables are traits such as knowledge, ability, etc. Latent traits affecting examinee performance are mathematically represented by the Cartesian vector $\vec{\theta}$ which is the set of all ordered number K-tuples $(\theta^1, \theta^2, \dots, \theta^K)$ in K-dimensional Euclidean space. K is taken to be the total number of psychological dimensions necessary to ensure a constant conditional distribution of the manifest variable over all populations of interest for a given value of $\vec{\theta}$. Thus, it can be said that the K-space is a complete latent space. When this condition is met, the assumption of local independence is satisfied. Local independence on a multiple choice test simply means that the probability of a response to a particular test item is not related to or affected by a response to any other test item except through the latent variables $\theta^1, \theta^2, \dots, \theta^K$. According to Lord and Novick (1968), "the assumption of local (conditional) independence, in its various forms, is the foundation of latent trait theory...".

The mathematics of latent trait theory is centered around the idea of the item characteristic function (ICF). The ICF describes the relationship between how individuals respond to a given test item and the extent to which they possess the particular trait measured by that test item. The ICF is defined as the regression of item score X_g on $\vec{\theta}$. This is similar to saying that the probability that an individual's response will assume a certain value (X_g) depends upon the amount of the trait measured by the item which the individual possesses. More precisely,

$$\text{Prob}(X_g = x_g) = f(\vec{\theta}), g = 1, 2, \dots, n$$

For dichotomous items x_g is equal to zero or one. It is customary to let:

$$\text{Prob}(X_g = 1 | \vec{\theta}) \equiv P_g(\vec{\theta}) \equiv P_g$$

and

$$\text{Prob}(X_g = 0 | \vec{\theta}) \equiv Q_g(\vec{\theta}) \equiv 1 - P_g$$

If $\vec{\theta}$ is unobservable, however, how can the form of $P_g(\vec{\theta})$ be determined?

There are several models which are used to estimate the shape of the ICF. The normal ogive model is probably the most familiar latent trait model presently available. On a multiple-choice test item, for which there is some finite probability of a correct response attributable to guessing, the three-parameter normal ogive model is appropriate. In addition to local independence, the three-parameter normal

ogive model depends upon the acceptance of the assumptions that the latent space is unidimensional and the item characteristic curve (ICC) representing the ICF has a normal ogive shape defined by the relation:

$$P_g(\vec{\theta}) = P_g(\theta) \equiv C_g + (1-C_g)(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{a_g(\theta-b_g)} \exp\left(-\frac{t^2}{2}\right) dt \quad (1)$$

where the three item parameters are:

a_g = item discriminating power

b_g = item difficulty level

c_g = probability of correct response by guessing

The assumption that the ICC has the above form can be satisfied if the following conditions are met.

Given that Γ_{ga} is a hypothetical trait of examinee a which affects his performance on test item g such that,

$$\Gamma_{ga} = \gamma_{ga} > \gamma_{gc} \Rightarrow X_{ga} = 1 \text{ and}$$

$$\Gamma_{ga} = \gamma_{ga} < \gamma_{gc} \Rightarrow X_{ga} = 0$$

where γ_{gc} is some constant value of Γ_g which characterizes item g , and, for the moment neglecting the affect of guessing on P_g , the following conditions are sufficient to establish the tenability of the normal ogive model:

1. $f(\Gamma_g|\theta) \sim N(m_1, m_2)$ where

$f(\Gamma_g|\theta) = \partial/\partial\theta \text{ Prob}(\Gamma_g|\theta)$ is the conditional frequency distribution of Γ for fixed θ , and where m_1 is the first moment about the origin and m_2 is second moment about the mean of this distribution. This is the continuous variable which is said to underlie the right-wrong dichotomous nature of the responses obtained from the item.

2. Γ_g is linearly related to θ in a stochastic manner. More precisely, the locus of m_1 for all values of θ is a straight line.
3. Homoscedasticity with respect to θ . This simply means that the values of m_2 do not depend upon the values of θ .

From the first condition above, the following relation holds:

$$f(\Gamma_g|\theta) = (2m_2\pi)^{-\frac{1}{2}} \exp\left[-\frac{(\gamma_g - m_1)^2}{2m_2}\right]$$

Integrating, we arrive at the cumulative normal distribution function

$$P_g = \text{Prob} (\Gamma_g > \gamma_{gc} | \theta) = \int_{\gamma_{gc}}^{\infty} f(\Gamma_g | \theta) d\gamma_g$$

It can be seen that

$$P_g = \text{Prob} \left[\left(\frac{\Gamma_g - m_1}{m_2} \right) > \left(\frac{\gamma_{gc} - m_1}{m_2} \right) \right]$$

$$= \int_{\frac{\gamma_{gc} - m_1}{m_2}}^{\infty} f \left[\left(\frac{\Gamma_g - m_1}{m_2} \right) | \theta \right] d \left(\frac{\Gamma_g - m_1}{m_2} \right)$$

where $f \left[\left(\frac{\Gamma_g - m_1}{m_2} \right) | \theta \right]$ is the standard normal frequency distribution.

In other words,

where $f(\Gamma_g | \theta) \sim N(m_1, m_2)$

now $f \left[\left(\frac{\Gamma_g - m_1}{m_2} \right) | \theta \right] \sim N(0, 1)$

Since the distribution is now standardized, we may write,

$$P_g = (2\pi)^{-\frac{1}{2}} \int_{\frac{\gamma_{gc} - m_1}{m_2}}^{\infty} \exp \left(-\frac{t^2}{2} \right) dt$$

and by symmetry,

$$P_g = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\frac{m_1 - \gamma_{gc}}{m_2}} \exp \left(-\frac{t^2}{2} \right) dt$$

Now since m_1 is a linear function of θ ,

$$m_1 = k_1 \theta$$

and since by condition 3 above, m_2 is constant

$$m_2 = k_2$$

then,

$$\frac{m_1 - \gamma_{gc}}{m_2} = \frac{k_1 \theta - \gamma_{gc}}{k_2}$$

$$= \frac{k_1}{k_2} \left(\theta - \frac{\gamma_{gc}}{k_1} \right)$$

let

$$k_3 = \frac{k_1}{k_2}$$

$$k_4 = \frac{1}{k_1}$$

then,

$$P_g = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{k_3(\theta - k_4\gamma_{gc})} \exp\left(-\frac{t^2}{2}\right) dt \quad (2)$$

Equation (2) is equivalent to equation (1), above, for the special case where $a_g = k_3$, $b_g = k_4\gamma_{gc}$, and $c_g = 0$. Since the effect of different specific values of a_g , b_g , and c_g does not change the shape of the ICC (it only changes its slope and orientation with respect to the coordinate axes) the above conditions are sufficient to establish that the form of the ICC is a normal ogive. It should be realized that the conditions stipulated above make no restriction on the frequency distribution of Γ_g or θ across the total population of examinees. Hence, the three-parameter normal ogive model holds even when θ is not normally distributed in a test population.

If a_g , b_g , and c_g are known for a particular item, its ICC is completely defined. These parameters do not vary from group to group as is the case with conventional item parameters. The parameter a_g is a measure of the slope of the ICC at its inflexion point. Since a steeply sloped ICC is sensitive to small differences in ability levels, a large value of a_g characterizes an item which discriminates well over a narrow range of ability levels. Similarly, a flat ICC has a small a_g and discriminates poorly over a relatively wide range of ability levels.

The center of the ability range over which a particular item discriminates is known as the difficulty level of the item, or b_g . The parameter b_g is the value of θ which corresponds to the abscissa of the inflexion point of the ICC. Therefore, a large value of b_g requires a high ability level in order to yield a high probability of a correct response. This indicates that b_g increases with the difficulty of the item.

The parameter c_g is the probability of a correct response to item g independent of θ . Even an examinee possessing no ability could theoretically answer some proportion of multiple-choice items correctly by mere chance. This proportion attributable to chance is designated c_g and determines the lower asymptote of the ICC. Methods of calculating estimates of the parameters a_g , b_g , and θ are presented in Appendix A.

SUMMARY

Item parameters (difficulty and discrimination indices) currently derived from conventional test items do not provide constant measures of the characteristics of the items they represent. Indeed, they are subject to wide variations from one test population to another. Similarly, examinee test scores typically defy comparison across test administrations due to the lack of a single, absolute score metric. A psychometric model for obtaining invariant item parameters along with a score metric which yields examinee ability assessments that lend themselves to comparisons between individuals in different test groups, has been presented in this paper. This model provides a comprehensive measurement system which is commensurate with the comprehensive nature of the personnel and training decisions required of an effective testing program.

APPENDIX A

METHODS OF CALCULATING NORMAL OGIVE TEST ITEM STATISTICS AND MAXIMUM LIKELIHOOD ABILITY ESTIMATES

The following proposed alternatives to conventional item parameters and scores were presented in the body of this paper:

Conventional Parameter or Score	Proposed Parameter or Score
VI (Validity Index)	a_g (Discrimination Power)
DI (Difficulty Index)	b_g (Difficulty Level)
Examinee Score on Test	θ (Examinee Ability Level)

We are confronted with the problem of how to obtain the proposed item parameters from the conventional ones. The following computational formulas provide a means of obtaining the desired conversions. (These conversions hold only to the extent that the examinees tested can be assumed to constitute a random sample from a population in which the ability θ is normally distributed.)

To obtain a good approximation of the discrimination power (a_g) of an item from its validity index (VI), the following formula can be used*:

$$a_g \approx \hat{\rho}'_g / \sqrt{1 - \hat{\rho}'_g{}^2}$$

where $\hat{\rho}'_g = \frac{VI \sqrt{DI'(1-DI')}}{(1-c_g) (2\pi)^{-1/2} \exp\left(\frac{\hat{\gamma}_g^2}{2}\right)}$, $\hat{\rho}'_g$ being an estimate of ρ'_g ,

$DI' = DI(1-c_g) + c_g$ and $\hat{\gamma}$ is determined from

$$DI = (2\pi)^{-1/2} \int_{\hat{\gamma}_g}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt, \hat{\gamma}_g \text{ being an estimate of } \gamma_g$$

To obtain a good approximation of the difficulty level (b_g) of an item from its difficulty index (DI), the following formula can be used.

$$b_g \approx \frac{-\hat{\gamma}_g}{\hat{\rho}'_g}$$

where $\hat{\rho}'_g$ and $\hat{\gamma}_g$ are calculated in the same manner shown above.

* Assuming that the VI used is the point biserial correlation between item and test area raw score, DI is the proportion of items answered correctly, and c_g is the "guessing" parameter.

Once approximations of a_g and b_g have been computed for each item on the test, they can be used to find the maximum likelihood estimate of an examinee's ability θ based upon the particular ensemble of item response scores characterizing that individual. This is done by maximizing the likelihood function specified below:

$$L = \prod_{g=1}^n P_g(\theta)^{X_g} Q_g(\theta)^{1-X_g}$$

where X_g is the score on item g and

$$P_g(\theta) = c_g + (1-c_g)(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{a_g(\theta-b_g)} \exp\left(-\frac{t^2}{2}\right) dt$$

and

$$Q_g(\theta) = 1 - P_g(\theta)$$

This involves using a_g and b_g for each item to calculate the values of P_g for the correct responses and Q_g for the incorrect responses in terms of θ . For a homogeneous set of test items (all measuring the same ability), a good statistical estimate of θ can be determined by finding the value of θ that maximizes L .

It should be emphasized that the accuracy of the calculations of this appendix depends upon the assumption of normality of θ in the population tested. Practical research by Lord (1952) and Indow and Samejima (1962, 1966) has shown that normal ogive estimates computed under the assumption of a normal distribution of θ yielded encouraging results for their data.

It is anticipated that relatively modest EDP support would be required for the maximum likelihood calculations discussed above. However, rather arduous methods exist for calculating a_g , b_g , and θ when θ is not distributed normally. A maximum likelihood technique which simultaneously solves for the three parameters can accomplish this objective. Due to the EDP expense of producing such solutions, and since the normality assumption is not considered to be unreasonable for most data, it is not recommended that such involved calculations be performed until the utility of the estimates described in this appendix is first investigated.

APPENDIX B

REFERENCES

- Indow, T., and Samejima, F., 1962. LIS measurement scale for non-verbal reasoning ability. Tokyo: Nihon-Bunka Kagakusha. (In Japanese)
- Indow, T., and Samejima, F., 1966. On the results obtained by the absolute scaling model and the Lord model in the field of intelligence. Yokohama: Psychological Laboratory, Hiyoshi Campus, Keio University. (In English)
- Lord, F. M., 1952. A theory of test scores. Psychometric Monograph, No. 7.
- Lord, F. M. and Novick, M. R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading Massachusetts.